INFORMATION AND CONTROL 11, 423–428 (1967)

Memory Increases Capacity

J. WOLFOWITZ*

Cornell University, Ithaca, New York and the Technion, Haifa, Israel

I. INTRODUCTION

During a conversation at the Royan (1965) meeting on coding theory Professors Norman Abramson and W. Wesley Peterson (of the University of Hawaii) posed to the writer, at his request, the following problem: To prove (what they said was part of the unproved "folklore" of information theory) that memory increases capacity. Some correspondence followed in an attempt by the writer to obtain a more precise formulation of the problem, in which the channel with memory would be comparable with the memoryless channel. The present formulation is due to the writer, who also profited from a conversation with Professor Frank L. Huband of Rice University; the latter's help in checking the manuscript is gratefully acknowledged. Thanks are due all these colleagues who, however, did not, until this paper was written, see the present formulation, much less approve it, and who are not responsible for any of its inadequacies. It seems to the author that, even if one rejects the present formulation of the statement in the title, the problem itself may have some interest.

Let $\{1, \cdots, a\}$ (respectively $\{1, \cdots, b\}$) be the input (resp. output) alphabet of two channels, 0 and M. Let $w(\cdot|\cdot|i)$, $i = 1, \cdots, c$, be c channel probability functions (c.p.f.'s); i.e., for each i, each $j = 1, \cdots, a$, and each $k = 1, \cdots, b$, $w(k|j|i) \geqq 0$, and

$$\sum_{k=1}^{b} w(k|j|i) = 1, \qquad \text{all } i \text{ and } j.$$

Let $p = (p_1, \cdots, p_c)$ be a row (probability) vector; i.e., all p's are $\geqq 0$, $p_1 + \cdots + p_c = 1$. Let $M^* = \{m(i, j)\}$ be a $c \times c$ stochastic

* Fellow of the John Simon Guggenheim Memorial Foundation. Research supported by the U.S. Office of Naval Research through contract Nonr 266(04)-(NR 047-005).

matrix whose states form one ergodic class (there are no transient states) and for which the vector p is the stationary measure. This means that $pM^* = p$, and that $(M^*)^s$ approaches a matrix all of whose rows are p, as $s \to \infty$.

The channel 0 operates as follows: At each letter the "state" of the channel is chosen at random, independently of all other choices, with probability p_i of choosing i, $i = 1, \cdots, c$. The choice is also independent of any letter sent (or received), and the result of the choice is unknown to both sender and receiver. If the ith c.p.f. is thus chosen and the letter j is being sent, then, no matter what letters were sent and received, the probability that the letter k will be received is $w(k \mid j \mid i)$. (The ith c.p.f. governs the transmission of the letter j.) The channel is obviously a memoryless channel with c.p.f. $w_0(\cdot \mid \cdot)$ given by

$$w_0(k \mid j) = \sum_{i=1}^{c} p_i w(k \mid j \mid i).$$

(See Wolfowitz (1964), Section 4.6, Channel I.) Call the capacity of this channel C_0.

Channel M differs from channel 0 only in the way the c.p.f. for transmitting each letter is chosen. As before, the choice is independent of all letters sent and received, and is unknown to sender and receiver. The c.p.f. for transmitting the first letter of each word (of length n, say,) is chosen according to the probability distribution p. Suppose i_1 is the index of the c.p.f. chosen to govern the transmission of the first letter. The probability that i_2 should be the index of the c.p.f. chosen to govern the transmission of the second letter is $m(i_1, i_2)$. Suppose i_2 is the index of the c.p.f. for the second letter. The probability that i_3 should be the index of the c.p.f. for the third letter is $m(i_2, i_3)$, etc., etc. Once the c.p.f. is chosen, the transmission of the letter being sent is governed by this c.p.f., and is independent of all letters previously sent and received.

Thus the successive states of channel M are chosen by the operation of a Markov chain. Let $u_0 = (x_1, \cdots, x_n)$ be any word (sequence) of length n (of n letters in the input alphabet) which is being sent (over either channel) and let $v(u_0) = (v_1(u_0), \cdots, v_n(u_0))$ be the chance word received. Let $v_0 = (y_1, \cdots, y_n)$ be any sequence of length n in the output alphabet. The symbol $P\{\cdot \mid M\}$ (resp, $P\{\cdot \mid 0\}$) will denote the probability of the relation in braces under channel M (resp., under channel 0). Thus

$$P\{v_i(u_0) = y_i \mid 0\} = w_0(y_i \mid x_i) \tag{1.1}$$

and

$$P\{v(u_0) = v_0 \mid 0\} = \prod_{i=1}^{n} w_0(y_i \mid x_i). \tag{1.2}$$

Since p is the stationary measure for M^* it follows that

$$P\{v_i(u_0) = y_i \mid M\} = w_0(y_i \mid x_i). \tag{1.3}$$

However, it is not in general true that, under M, $v_1(u_0), \cdots, v_n(u_0)$ are independently distributed. Thus channel M has a "memory", while channel 0 is memoryless (see (1.2)). On the other hand, because of (1.1) and (1.3) the two channels are directly comparable.

Let C_0 (resp. C_m) be the capacity of channel 0 (resp., channel M). The result of the title of this paper is

THEOREM 1. $C_m \geqq C_0$.

This will be proved in Section 3. We will also obtain C_m (Theorem 2 below). The value of C_0 is, of course, well known (e.g., Wolfowitz (1964), Section 3.1).

2. AN AUXILIARY RESULT

Let $H(Z_1)$ denote the entropy of the chance variable Z_1 (see Wolfowitz (1964), Chapter 2) and let $H(Z_1 \mid Z_2)$ denote the entropy of Z_1, given the chance variable Z_2. *In this section only* let $X_1, \cdots, X_n$ be independent chance variables with values in the input alphabet and a common distribution which we shall not need to specify. Let $Y_1, \cdots, Y_n$ be chance variables (to be described in a moment) with values in the output alphabet. Write, for brevity,

$$X^{(n)} = (X_1, \cdots, X_n), \qquad Y^{(n)} = (Y_1, \cdots, Y_n).$$

Then we define

$$Y^{(n)} \equiv v(X^{(n)}).$$

Thus the distribution of $Y^{(n)}$ and of $(X^{(n)}, Y^{(n)})$ depends upon the channel, while the distribution of $X^{(n)}$ does not. We shall write H_m for entropy when the distribution is determined by channel M (for example, $H_m(Y^{(n)})$, $H_m(Y^{(n)} \mid X^{(n)})$) and H_0 for entropy when the distribution is determined by channel 0 (for example, $H_0(Y^{(n)})$, $H_0(Y^{(n)} \mid X^{(n)})$). Since the distribution of $X^{(n)}$ does not depend on the channel we may, for example, write $H(X^{(n)})$, $H_m(X^{(n)})$, or $H_0(X^{(n)})$ at pleasure. We will now

prove that

$$H(X^{(n)}) - H_m(X^{(n)} \mid Y^{(n)}) \geqq H(X^{(n)}) - H_0(X^{(n)} \mid Y^{(n)}). \quad (2.1)$$

This follows immediately from

$$\begin{aligned} H_m(X^{(n)} \mid Y^{(n)}) &\leqq \sum_{i=1}^{n} H_m(X_i \mid Y^{(n)}) \\ &\leqq \sum_{i=1}^{n} H_m(X_i \mid Y_i) = \sum_{i=1}^{n} H_0(X_i \mid Y_i) \qquad (2.2) \\ &= H_0(X^{(n)} \mid Y^{(n)}). \end{aligned}$$

3. THE CAPACITY OF CHANNEL M

We now drop the requirement that $X_1, \cdots, X_n$ be independently and identically distributed, and denote their joint distribution by Q_n', the notation used in Wolfowitz (1964). We shall now prove

THEOREM 2.[1] *The capacity of channel M is given by*

$$C_m = \lim_{k \to \infty} \left\{ \frac{1}{k} \sup_{Q'_k} [H(X^{(k)}) - H_m(X^{(k)} \mid Y^{(k)})] \right\}. \quad (3.1)$$

The proof of Theorem 2 will lean heavily on the methods of Wolfowitz (1964); where a technique from Wolfowitz (1964) is used the argument will be sketched and the reader referred to Wolfowitz (1964) for details. The existence of the limit in (3.1) will be proved at once. Let $W(k)$ be the quantity in curly braces in (3.1).

Let $\epsilon > 0$ and $0 < \lambda < 1$ be arbitrary but fixed. We shall show the following: Let ℓ be sufficiently large, and then n, depending on ℓ, sufficiently large.

(3.2) For all such n there exists a code

$$(n, \exp_2 \{n(W(\ell) - \epsilon)\}, \lambda)$$

[1] (Added June 30, 1967, while this manuscript was at the editor's.) A very special case of channel M is studied by Bruce D. Fritchman (A binary channel characterization using partitioned Markov chains. *IEEE Trans. Inform. Theory*, **IT13**, Number 2, April, 1967, 221–227) for different purposes. For this particular case Theorem 2 of the present paper is stated, and references are cited which are supposed to prove it. The references are irrelevant because they deal with entirely different channels.

for channel M, and

(3.3) For all such n there does not exist a code

$$(n, \exp_2 \{n(W(\ell) + \epsilon)\}, \lambda)$$

for channel M.

Since ϵ was arbitrary, and since (3.2) and (3.3) are valid with different large ℓ, it follows that if the limit in (3.1) did not exist, (3.2) and (3.3) would lead to a contradiction. Hence the limit in (3.1) exists and it remains only to prove (3.2) and (3.3).

To prove (3.2) we use the ideas of Wolfowitz (1964), (Sections 5.3 and 6.7). The word of length n is made up of many blocks of length $(\ell + d)$. The last d letters of each block are "wasted" and not used for the message. The ratio d/ℓ is very small, so that only a very small fraction of the letters sent is wasted; this fraction is taken up in the ϵ of (3.2). The number d is large, so large that the state of the channel after the d wasted letters have been transmitted, i.e., the c.p.f. which governs the transmission of the first letter of the next block of $(\ell + d)$ letters, is selected with a probability distribution very close to p. (Recall that $(M^*)^s$ approaches a matrix all of whose rows are p, as $s \to \infty$.) Thus each block of $(\ell + d)$ letters (or rather, its first ℓ letters which actually carry the message) starts anew, as it were, and the power of the memory between the blocks of ℓ letters which carry the message is small. The desired result now follows from Wolfowitz (1964), (6.7.3).

The proof of (3.3) is scarcely different from that of Wolfowitz (1964), (6.7.2) and its antecedent Lemma 6.6.5, making use of some of the arguments of the preceding paragraph. This completes the proof of Theorem 2.

In Wolfowitz (1964) we have stressed the importance of a constructive description of the capacity, i.e., the importance of being able to compute the capacity of a channel to within any desired accuracy. By methods similar to those used to prove Theorem 2, one can obtain a (necessary) bound on the rapidity of approach of $W(k)$ to C_m. We shall content ourselves with stating the results without proof.

Let δ be the function defined in Wolfowitz (1964), (6.6.2), $c(\cdot)$ the function defined in Wolfowitz (1964), (6.6.15), and $g = \max(a, b)$. Then one can show without difficulty that, for any (integral) d and ℓ,

$$(3.4) \quad |C_m - W(k)| \leqq \frac{d \log g}{d + k} + 8g(\log_2 g)\delta((M^*)^d) + 4gc(\delta(M^*)^d).$$

This bound can easily be improved.

We shall now prove Theorem 1. From (2.1) and (3.1) we have that

$$W(k) \geqq C_0 . \tag{3.5}$$

Hence $C_m \geqq C_0$, as was to be proved.

4. GENERALIZATION

Only the ergodic property of the Markov matrix M^* was used above, as pointed out to me by Mr. Samuel Friedland of the Israel Technion. Thus we have already proved the following:

THEOREM 3. *Let M' be any channel (with the same alphabets as M) such that the power of the memory between blocks of letters separated by d letters approaches zero as $d \to \infty$, uniformly in the blocks, and such that, for every positive integer i, every input word u_0 and every output letter t,*

$$P\{v_i(u_0) = t \mid M'\} = P\{v_i(u_0) = t \mid 0\}.$$

Then the capacity of M' is not less than that of 0, and is given by an expression which corresponds to the right member of (3.1).

RECEIVED: April 3, 1967

REFERENCE

WOLFOWITZ, J., (1964). "Coding Theorems of Information Theory." 2nd ed., Springer, Berlin and New York.